Food and Agriculture Organization of the United Nations

AFRICAN COMMISSION ON AGRICULTURAL STATISTICS

Twenty-Seventh Session

15 – 18 November 2021, Virtual Host – Dakar, Senegal

AGENDA ITEM 5

GUIDELINES ON DATA DISAGGREGATION FOR SDG INDICATORS USING SURVEY DATA

Clara Aida Khalil
Statistician
Office of the Chief Statistician - FAO

Data disaggregation and the SDGs

- With the adoption of the 2030 Agenda for Sustainable Development, Member States have pledged to **leave no one behind (LNOB)** and reach the furthest behind first: Need for more **disaggregated data** than currently available in most countries.

An overarching **principle of data disaggregation** is at the core of the SDG Monitoring Framework:

> *"SDG Indicators should be disaggregated, where relevant, by income, sex, age, race, ethnicity, migratory status, disability and geographic location, or other characteristics in accordance with the Fundamental Principles of Official Statistics."*

Data disaggregation and the SDGs (2)

The IAEG-SDGs formed a Working Group on Data Disaggregation to strengthen national capacities and develop the necessary statistical standards and tools to produce disaggregated data. This led to:

- The development of a **minimum disaggregation set** and the compilation of **categories and dimensions of data disaggregation** currently in place and planned by custodian agencies.
- A comprehensive summary of **disaggregation standards and classifications** for all SDG Indicators.
- A compilation of **policy priorities** by disaggregation dimension.
- A compilation of **methods and tools** for data disaggregation.

Resources accessible from: https://unstats.un.org/sdgs/iaeg-sdgs/disaggregation/

Guidelines on data disaggregation for SDG indicators

As a member of the **WG on data disaggregation** and the **TA on SAE**, the FAO has developed «*Guidelines on data disaggregation for SDG Indicators using survey data*».

Main objectives of the Publication:

- Offer methodological and practical guidance for the production of **direct** and **indirect** disaggregated estimates of SDG Indicators.
- Provide tools to assess estimates accuracy and present strategies for data integration, including **small area estimation** (SAE) methods

Link to the Guidelines: http://www.fao.org/documents/card/en/c/cb3253en/

Relevance of the guidelines

- Approximatively 1/3 of the Global SDG Indicators can be computed using survey data.
- 6 out of 21 SDG Indicators under FAO custodianship can rely on data from household and/or agricultural surveys.

ISSUE ADDRESSED: The use of traditional sampling techniques imposes limitations on the production of disaggregated data and reliable estimates for small sub-populations. Innovative techniques that could address some of these issues are far from being mainstreamed in National Statistical Offices.

Data disaggregation with sample surveys

The guidelines in a nutshell:

- **Direct estimates of an indicator for a given sub-population:** based only on sample information from the sub-population itself. **Two main issues:**
  - Sampling size often not large enough to guarantee reliable estimates for small domains;
  - Possibility of having non sampled sub-domains.
- These issues can be addressed:
  - **At the design stage:** adopting sampling designs that guarantee an observed set of sampling units for every sub-population for which disaggregated data must be produced.
  - **At the analysis stage:** producing indirect estimates, coping with the little information available for "small areas" by borrowing strength from other data sources.

Addressing data disaggregation at sampling design stage

In order to produce **direct disaggregated estimates**, sampling strategies should ensure the presence of a sufficient number of sampling units in each disaggregation domain.

Moreover, having sampling units in each disaggregation domain is also **beneficial to indirect estimation**: reduction of model bias and variance

- **Straightforward** when the number of units belonging to a given sub-population can be determined from the sampling frame. In these cases, the main issue is the selection of the degree of oversampling to apply.

- **More complex** when members of sub-populations are not known in advance from the available sampling frame.

Addressing data disaggregation at sampling design stage (2)

Mainstreamed and innovative approaches to ensure sufficient sample size for each disaggregation domain are illustrated and discussed (with their pros and cons), including:

- Oversampling
- Deeper stratification
- Multiphase sampling with screening of respondents

→ **TRADITIONAL**

- Marginal stratification designs
- Indirect sampling

→ **INNOVATIVE**
Ensure a sufficient sample for each disaggregation domain without increasing the overall sample size

Suitable software packages are suggested

Addressing data disaggregation at the analysis stage

- The guidelines illustrate **alternative sampling strategies** for direct domain sampling estimation
- Most common **direct and indirect model-assisted domain estimators** are discussed, introducing their context of usability.
- **Small Area Estimation** (SAE) techniques are illustrated:
  - Presenting the process flow for their **implementation**;
  - Providing an overview of main **unit-level** and **area-level** approaches;
  - Indicating main references on the topic;
  - Giving tools to assess the quality of small area estimates.

Addressing data disaggregation at the analysis stage (2)

Projection estimator (Kim and Rao, 2012) is introduced, discussed and experimented on actual survey data.

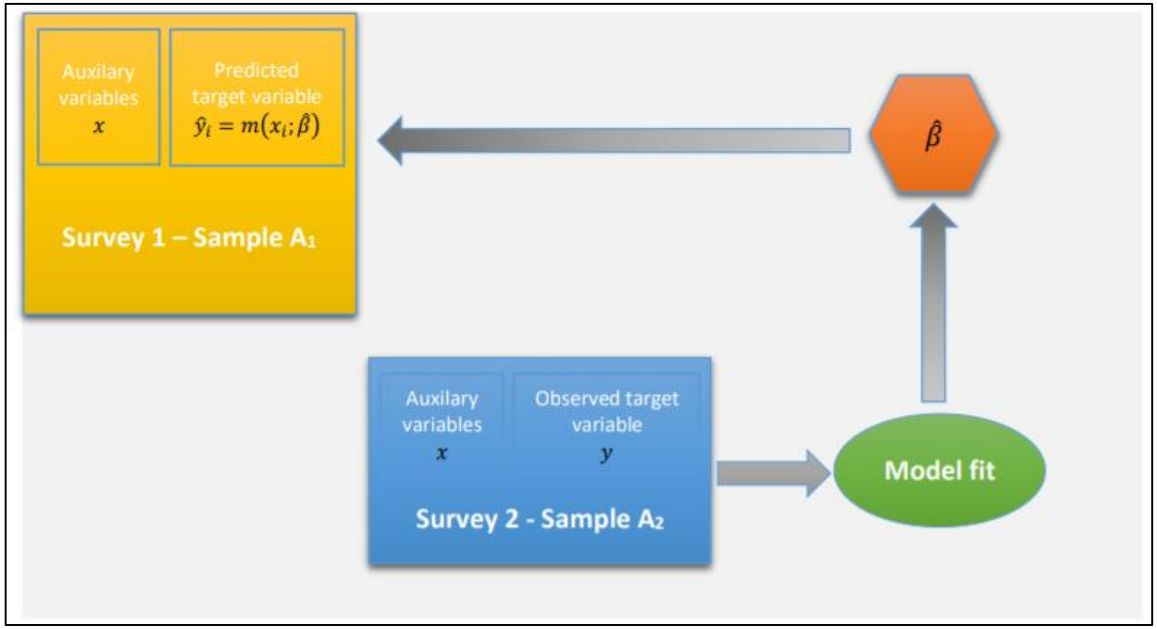
It allows producing disaggregated indicators by the **joint use of two sample surveys:**

- **The first survey**, is characterized by a large sample A_1, but only collects auxiliary information or variables of general use (e.g. socio-economic variables);
- **The second survey** has a smaller sample A_2 but collects information on the target variable y, along with the same set of auxiliary variables available from A_1.

Addressing data disaggregation at the analysis stage (3)

The total of variable y in the disaggregation domain d can be estimated as

$$\hat{Y}_{PR,d} = \sum_{i \in A_1} w_{i1} m(x_i; \hat{\beta}) \gamma_{id}$$



Assessing estimates accuracy

The publication emphasizes the importance of **estimating** and **disseminating** accuracy measures:

- To enable users assessing the fitness for use of an estimate.
- To build public trust in data and their use.

Methods and Tools to assess the accuracy of direct estimates are provided:

- **Sampling variance estimation:** to be used when computation of indicators is based on the inferential properties of repeated sampling schemes.
- **Model Variance:** suitable when estimation relies **only** on models using auxiliary variables.
- **Global Variance:** when model-based approaches are used jointly with inference based on the sampling design.

A practical application based on SDG Indicator 2.1.2

The approach has been adopted to produce disaggregated estimates of **SDG Indicator 2.1.2** on the Prevalence of Moderate and Severe Food Insecurity based on the Food Insecurity Experience Scale (FIES).

Objective: Estimate Indicator 2.1.2 by sex, rural/urban, age class, and income quintile.

Two data sources:

- Malawi's Fourth Integrated Household Survey (IHS4) 2016-17
- Malawi FIES survey module collected through the Gallup World Poll – 2016

Results are presented along with accuracy measures of indirect estimates.

A practical application based on SDG Indicator 2.1.2 (2)

The practical implementation of the case study can be summarized with following steps:

I. **Identifying and recoding auxiliary variables.** The implementation of the *Projection estimator* requires the availability of the same set of auxiliary variables in the two surveys to be integrated, with common structure and definitions.

II. **Definition of the function $m()$ and estimation of projection parameters.** The selection of the functional form for m relies heavily on the type of variable y considered (e.g. scale, nominal, dichotomous).

III. **Computation of synthetic values.** Using the estimated projection parameter, the synthetic values of the variable of interest in the large dataset are computed.

IV. **Assessment of estimates accuracy.** Estimation of the variance, coefficient of variation and confidence intervals for the projected indirect estimates.

A practical application based on SDG Indicator 2.1.2 (3)

Some results:

		Moderate or severe food insecurity			
		Prob.ms	CV (%)	Lower_CI	Upper_CI
IHS4	Total	0,91	1,2	0,89	0,93
GWP		0,91	1,3	0,89	0,93
IHS4	Female	0,91	1,4	0,88	0,93
GWP		0,90	1,5	0,89	0,94
IHS4	Male	0,91	1,9	0,87	0,94
GWP		0,91	2,0	0,87	0,94
IHS4	Rural	0,93	1,2	0,90	0,95
GWP		0,92	1,3	0,90	0,94
IHS4	Urban	0,81	5,7	0,73	0,92
GWP		0,82	5,9	0,74	0,93
IHS4	15-24	0,91	2,0	0,87	0,94
GWP		0,89	2,1	0,85	0,93
IHS4	25-49	0.91	1,6	0,88	0,93
GWP		0.92	1,6	0,89	0,95
IHS4	50-64	0,87	3,6	0,82	0,94
GWP		0,90	3,5	0,84	0,96
IHS4	65+	0,97	1,6	0,94	1
GWP		0,98	1,7	0,95	1
IHS4	Inc_1	0,96	1,5	0,94	0,99
GWP		0,97	1,5	0,94	1
IHS4	Inc_2	0,96	1,5	0,93	0,99
GWP		0,96	1,6	0,93	0,99
IHS4	Inc_3	0,97	1,1	0,95	0,99
GWP		0,97	1,1	0,95	0,99
IHS4	Inc_4	0,89	3,6	0,82	0,95
GWP		0,88	3,7	0,82	0,94
IHS4	Inc_5	0,74	3,8	0,68	0,80
GWP		0,76	3,8	0,71	0,82

		Severe food insecurity			
		Prob.s	CV (%)	Lower_CI	Upper_CI
IHS4	Total	0,73	2,4	0,67	0,75
GWP		0,71	2,8	0,67	0,75
IHS4	Female	0,75	2,8	0,71	0,80
GWP		0,75	3,1	0,71	0,80
IHS4	Male	0,70	3,6	0,65	0,75
GWP		0,67	4,2	0,61	0,73
IHS4	Rural	0,75	2,4	0,72	0,79
GWP		0,72	2,9	0,68	0,76
IHS4	Urban	0,63	9,5	0,51	0,75
GWP		0,63	9,2	0,52	0,75
IHS4	15-24	0,72	3,8	0,66	0,77
GWP		0,67	4,5	0,61	0,73
IHS4	25-49	0,71	3,2	0,67	0,76
GWP		0,72	3,6	0,67	0,77
IHS4	50-64	0,74	6,9	0,64	0,84
GWP		0,75	7,1	0,65	0,86
IHS4	65+	0,86	5,4	0,75	0,96
GWP		0,87	5,8	0,78	0,98
IHS4	Inc_1	0,88	3,3	0,83	0,94
GWP		0,88	3,4	0,83	0,94
IHS4	Inc_2	0,82	3,8	0,75	0,88
GWP		0,81	4,2	0,75	0,88
IHS4	Inc_3	0,77	5,3	0,69	0,85
GWP		0,75	5,3	0,67	0,83
IHS4	Inc_4	0,68	6,0	0,60	0,76
GWP		0,64	6,5	0,56	0,72
IHS4	Inc_5	0,48	8,4	0,40	0,56
GWP		0,48	8,4	0,40	0,56

THANK YOU!

ClaraAida.Khalil@fao.org